THE EXTERNAL VALIDITY IN ECONOMIC EXPERIMENTS

By Mar Miquel

Presented to Professor Karl Schlag

Universitat Pompeu Fabra

Barcelona, December 6th 2012

Abstract

One of the most important questions regarding experimental economics is the external validity of laboratory experiments. This paper goes through a study that tests the generalizability of a Dictator Game as a laboratory analogue for a naturally occurring decision-making context of teacher absenteeism. Because lab and naturally-occurring environments systematically differ we then discuss other factors that might strongly affect the choices that individuals make. We conclude that the dichotomy drawn between lab experiments and data collected from natural settings is a false one. A combination of the two would provide deeper and better insights than either separately.

Introduction

Nearly 400 years ago, Galileo tried to test his theory of acceleration timing balls as they rolled down an inclined plane. He died not knowing that that was the first recorded laboratory experiment ever performed in history. Since then, lab experiments have been the key element when applying the scientific method in any physical science. According to Feynman, the principle and definition of science is that the test of all knowledge is experiment; Experiment is the sole judge of scientific "truth".

Traditionally, economics was thought to be a non-experimental science that had to rely on observations from real-world economies rather than controlled laboratory experiments. Economics differs from other physical sciences such as biology or physics because it is studying human behavior, and it cannot easily control nor predict a lot of important factors, and it was thought that it needed significantly large groups of sample to reach relevant conclusions.

This was viewed as a huge obstacle to continue the development of economics as a science. However, the creation and establishment of experimental economics, a growing research field, has exponentially challenged this perception. Under controlled laboratory conditions, experimentalists study human behavior in situations that, in simplified and pure forms, mirror those found in "real" markets and other forms of economic interaction. The attractiveness of experiments is that, in principle, they provide *ceteris paribus* observations of motivated individual economic agents, which are otherwise extremely difficult to obtain using conventional econometric techniques.

The extent to which the results of such experiments can be generalized to market situations or to the "real world" is still under debate, and this is exactly this paper's aim: to determine if we can extrapolate the results from the lab to the real world in order to predict human behavior.

One may wonder why this question is crucial in Economics but it is not so important in the other physical sciences. The answer is that in Experimental Economics the subject of study are human behavior, as said before, and their choice-making ability, thus the extent to which researchers can generalize their conclusions or extrapolate them to the real world is a cornerstone of any research.

Internal vs. External validity

Before delving into the critical assumption we were just talking about, let us introduce the concept of **internal validity**. The internal validity of an experiment is the approximate truth about inferences regarding cause-effect or causal relationships; it refers to the ability to draw confident causal conclusions from the research. An experiment that is internally valid will yield results that are robust and replicable. An experimental result is *internally* valid if the experimenter attributes the production of an effect B to the factor A, and A really is the cause of B in the experimental set- up E.

The critical maintained assumption underlying many laboratory experiments is that the insights gained in the lab can be extrapolated to the world beyond, a principle that Levitt and List (2006) denote as *generalizability*. In this paper, we are going to use the term “external validity” as a synonym.

External validity, a term that was first used by Campbell and Stanley back in 1963, asks the question of generalizability: to what population, settings, treatment variables and measurement variables can this effect be generalized? This way, it refers to the possibility of *generalizing* the conclusions to situations that prompted the research. There is an obvious tension between the two. Where internal validity often requires abstraction and simplification to make the research more tractable, these concessions are made at the cost of decreasing external validity.

The most often heard comment on experimental studies, and the motivation of this paper/research is that the experiments may not reflect the ‘real world’ and therefore they may or may not teach us much about economics nor we can extrapolate conclusions to the *real* world.

However, not all experiments are designed to test formal theory. Charlie Plott, for example, argues that experiments do not need to be realistic so long as they closely implement the theory being tested[1]. Hence, this approach primarily sees experiments

[1] His argument is that experiments should be used along with theory: “Thus theory (...) serves importantly to simplify the experimental process. The more that accepted theory can be invoked, the less that the experimental process needs to “mirror” the natural analog.” (Plott, 1982, p.1521)

as a method for testing and comparing theories. Some of them are designed to produce stylized facts, or exhibits, about the way humans behave that may explain certain observed phenomena and could usefully become the foci of future theories. Others, though, are designed to explore policy-relevant hypothesis that, even though they are not founded on formal theory, are worthy of careful investigation. In this endeavors external validity is considerably more important.

The factors that compromise the external validity of economic experiments have been characterized in a variety of ways. In essence, three types of factor are a cause of concern: omissions, contaminations and the artificiality of alteration.

Omissions are factors in the naturally occurring decision-making context or system that are excluded from the experiment. That factors can be excluded or held constant in experiments when, in nature they are present and vary is, of course the great strength of experiments. However, especially where factors that are omitted or held constant in the lab may, in nature, interact with the factors under investigation in the lab, it is reasonable to question external validity.

The problem of **contamination** refers to the factors that are present in the lab but not in the nature. These include "Hawthorne effects", "demand effects" and "experimenter effects".

Finally, the **artificiality** of the setting refers to the fact that if the laboratory, incentives and the ambient are not close enough to the outside-the-laboratory situation it is intended to study, the loss of external validity may be significant. Indeed, it is obvious that it is absolutely necessary to interpret the word 'sufficiently' and this will vary substantially, depending on the goal and/or objective of the experiment.

Its main critique, named "the artificiality of alteration critique", argues that the relational spaces in the lab and in nature are fundamentally different. This way, while an experiment could be in principle be described as an analogue for the decision-making context it is designed to elucidate, it will always be artificial because it will never be exactly the same as reality, where natural decision-making occurs.

In this paper we will first go through a real study, developed by Abigail Barr and Andrew Zeitlin, and then we will compare it with one of the most known games in

Game Theory, the Dictator Game, and we will try to test the external validity of the experiment by comparing the results obtained in the field and in the economics lab.

Situation

In a recent study, Chaudhury et al. (2006) conducted unannounced spot checks in primary schools in Uganda and found 27% primary school teachers absent from their places of work. Ugandan local governments tend to be extremely under-resourced and, especially in the case of remote primary schools, incapable of monitor and discipline poorly performing teachers and school managers. Even the most extreme examples or practices of teacher absenteeism tend to be punished only by transfer to an even more remote school (to which they may never turn up, while continuing to earn a salary, so not real punishment happens).

Ugandan Study: basic information

The **focus** of the experiment is that teacher absenteeism is a significant problem in publicly funded schools throughout the developing world. The **aim** of the experiment is to establish whether a Dictator Game can be used simultaneously:

a) As a baseline in a series of laboratory experiments designed to investigate what would happen if Ugandan school management committees (SMCs) were empowered to hold teachers to account

b) To generate a measure of teacher's intrinsic motivations.

The **specific** of the experiment is to test and investigate the external validity of the DG.

The dictator game

In the Dictator Game (DG), the first player, *the proposer* (or Dictator), determines an allocation (split) of some endowment (such a cash prize). The second player, *the responder*, simply receives the remainder of the endowment left by the proposer. The responder's role is entirely passive: he/she has no strategic input into the outcome of the game.

In this particular case, the DG design preludes, to the extent possible, giving in the hope of reciprocation and giving to avoid punishment. Thus, it is well suited to the function of generating a measure of teacher's intrinsic or internal motivations. Plus, it appeared to be a good match for the status quo in Ugandan schools.

A characterization of the status quo

Teachers sell a contracted amount of time to the government each month. The government gives back this time and sends them off to remote communities to use the time to teach. But since the teachers are not monitored, their contracts are not enforced; hence they are free to choose how much time to allocate to teaching and how much to themselves. This looks like a **Dictator Game**: the teachers are the dictators; the communities are the recipients; the endowment is the teacher's time; and the size of the stake is specified in the contract.

Economics lab experiment: design and performance

They constructed one lab in each of 100 Ugandan primary schools. Then, they performed one session in each involving five teachers in the role of dictator, five parents of pupils in the role of recipient and five SMC members (present and paid but passive in the DG). Teachers and parents were randomly and anonymously paired. They played one round of one-shot DG, with a stake of 5,000 Ugandan Shillings (just under $3.00).

Each experimental session was conducted by four field researchers, using a large-enough classroom that could seat fifteen people and three decision-making stations (three other rooms or classrooms) located outside that main classroom, far enough to ensure complete privacy for one-on-one interviews.

The first thing researchers would do when the subjects arrived is to register each of the subjects and give each one of them a badge bearing a number and either an orange (for parents), green (for teachers) or blue (for SMC members, including the head master) figure. After that, each subject was invited to sit in the area in the classroom assigned to his or her badge color.

Once the subjects had been taught the game[2], the teachers were called to one-on-one meetings in one of the three decision-making stations with field researchers. They were taken through the game again, and were then asked to represent their chosen allocation by dividing the real-money stake between the green figure (representing themselves) and the orange figure (representing the parent they had been anonymously paired with) on the table in front of them. The stake in the DG was 5,000 Ugandan Shillings, presented in the form of ten 500 Ugandan Shilling coins, so all they had to do was to distribute their amount of coins trying to show their chosen allocation. Once all the teachers had made their decisions and returned to the classroom, a second game (a third party punishment game) was presented and played.

Finally, in order to test the external validity of the DG by correlation, it is needed an observational or survey measure that, under the assumption of external validity, will be correlated with the allocations made by the dictators in the Dictator Game.

In order to try to assess the potential correlation to the DG allocations that the teachers made to the parents, the authors of the study recorded the proportion of contracted time that each teacher allocated to teaching during the previous month to the survey and experiment. In the surveys each teacher was asked how many days they were absent from the classroom in the month prior to the survey. All the responses given by the teachers regarding this particular question could be easily checked if true or valid, as researchers were able to check each school's records. In the analysis, to minimize the effects of absenteeism being underreported (strategically on the part of teachers and resulting from poor record-keeping on the part of schools), the maximum of the two measures of absenteeism is used in conjunction with the assumption that teachers are contracted to work 20 days a month to calculate the proportion of contracted time that each teacher allocates to the community they have been sent to serve.

Theoretical framework

From here, we are going to use Barr and Zeitlin's theoretical framework in order to draw conclusions later on. The model they used and that is used in this paper also is a

[2] This would be done orally, as the researchers knew some subjects belonging to the parents and SMC groups would be illiterate.

simple theoretical model that yields a testable prediction about how teacher's allocations in the DG and contracted-time allocations in nature are related under the assumption that a single preference parameter is driving behavior in both contexts.

$$U_{it}^{k} = X_{it}^{k} - W^{k}\alpha_{t}\sum_{j=i,-i}\left(X_{jt}^{k} - Y_{j}^{k}\right)^{2}$$

Taking into account that X_{it}^{k} is the allocation by teacher *t* to *j* (*i* = self, *-i* = other) in context *k* (*S* = contracted time, *L* = Dictator Game), Y_{j}^{k} is the reference point allocation to *j* in context *k*, α_{t} is the preference to adhere to reference point allocations (RPAs) and W^{k} is the weight applied to reference point in context *k*.

The previous utility function presents something new with respect to the classical utility functions widely used in behavioral economics: the context-specific weights attached to the common-across-contexts preference parameter, which correspond to variations in the relevance of a given preference across different contexts and that's the reason why they could be interpreted as a way of formally capturing one dimension of the artificiality of alteration. In the current case, the contract relating to teachers' time allocations may strengthen the relevance of the preference in the natural as compared to the lab context.

Alternatively, the fact that the Dictator Game played in the economic lab created for the occasion is played with a windfall could strengthen the relevance of a preference to share-and-share-alike in the lab as compared to the natural context.

Once maximized the utility function wrote above subject to the normalizing constraints that $X_{it}^{k} + X_{-it}^{k} = 1$ and $Y_{i}^{k} + Y_{-i}^{k} = 1$ we find teachers *t*'s optimal allocation to the community in each of the contexts to be:

$$X_{-it}^{S*} = Y_{-i}^{S} - \frac{1}{2W^{S}\alpha_{t}} \quad \text{and} \quad X_{-it}^{L*} = Y_{-i}^{L} - \frac{1}{2W^{L}\alpha_{t}}$$

From here, we can state that the amount of time teachers allocate to teaching (teacher's allocations to the community) and their allocations of money in the DG are linearly related. Holding Y_{-i}^{S}, Y_{-i}^{L}, W^{S} and W^{L} constant and varying α_{t} we can write the linear relationship mentioned before:

$$X^{S*}_{-it} = \beta_0 + \beta_1 X^{L*}_{-it} \quad \text{where} \quad \beta_0 = Y^S_{-i} - \frac{W^L}{W^S} Y^L_{-i} \quad \text{and} \quad \beta_1 = \frac{W^L}{W^{S*}}$$

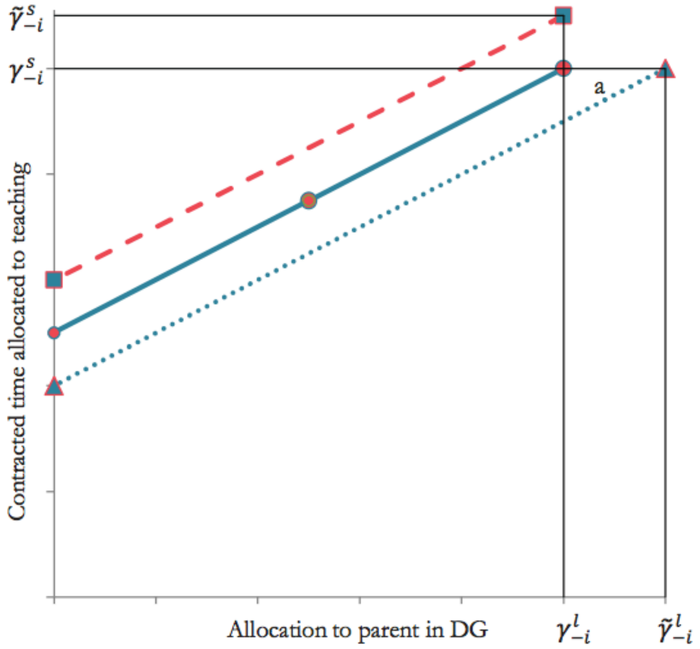

Figure 1: Predicted relationship between teachers' proportional allocations to parents in the Dictator Game and the proportion of contracted time allocated to teaching[3].

To test the external validity of the Dictator Game ran in Uganda what researchers did was to check the correlation between the time teachers allocated to teaching (this information was obtained through the surveys we mentioned before) and the allocations they did to parents in the DG. The solid blue line in **Figure 1** plots the relationship. Point ***a*** lies at $[Y^L_{-i},\ Y^S_{-i}]$. No teacher will be located above and to the right of point ***a***.

That means that if Reference Point Allocations (RPA, in this model α_t) and preference weights are common enough across all teachers and the preference to stick to a context-specific RPA is stable within teachers across context while varying across teachers, we will then have a linear relationship between allocations in the two contexts: in the field study and in the laboratory.

[3] Barr, A. and Zeitlin, A., 2010: "Dictator Games in the Lab and in nature: Evidence of external validity from Ugandan primary schools"

In **figure 1** (solid blue line) one can easily see that there is a positive correlation between the two variables we were just discussing, which means that Barr and Zeitlin found evidence that proved that the more money teachers allocated to parents in the Dictator Game, the more time they spent teaching (or what it's the same: the less they were absent from school).

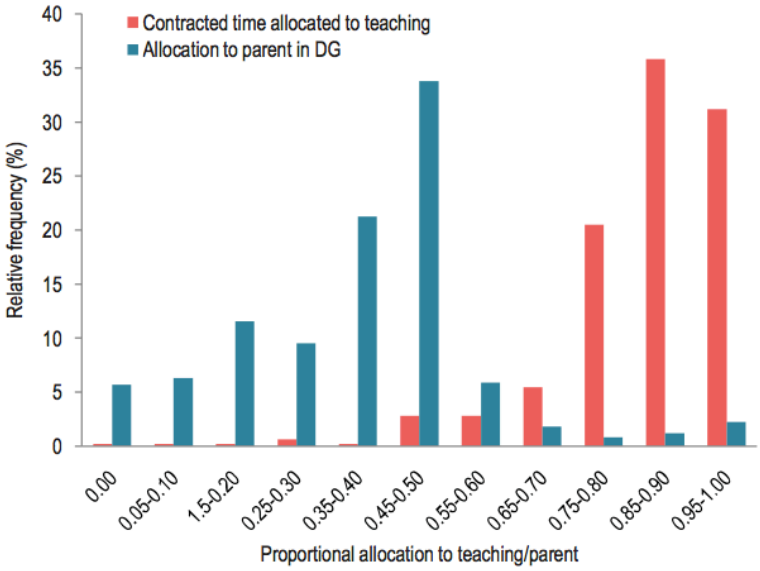

Figure 2: Distribution of contracted time allocations to teaching and allocations to parents in Dictator Game[4]

Figure 2 presents the distribution of the proportion of contracted-time that the teachers allocate to teaching (pink) and the proportional allocations teachers chose to make to parents during the Dictator Game (blue). One can easily see that both distributions present obvious differences: the DG allocations show a strong mode at 0.50 while time allocations display a strong mode at 0.85 to 0.90 (meaning 17-18 days in a 20 working day/month) and have a smaller variance than the DG allocations and show signs of truncation at 1.00. In the absence of a theory, the differences in the histograms may have led us to conclude that the DG is not a good match to the contracted-time allocation decisions being made by the teachers. However, the differences are consistent with the RPA and preference weight differing across the

[4] Barr, A. and Zeitlin, A., 2010: "Dictator Games in the Lab and in nature: Evidence of external validity from Ugandan primary schools"

two contexts and do not preclude a correlation driven by within-teacher stability and cross-teacher variations in the preference parameter, α_t.

The external validity of the lab-type DG

P-value of Spearman’s rho	0.013
P-value of pairwise correlation	0.042
P-value of linear bivariate regression	0.042
P-value of linear bivariate regression, clustering by school	0.063
P-value on F-test of joint sig. of teachers fixed effects	0.039

Table 1: statistical significance of relationship between proportion of contracted time allocated to teaching and proportional allocation to parent in Dictator Game (n=487)[5]

The point we were just discussing can be seen in **Table 1**. The first four P-values test Prediction 1 directly, because they relate to the existence of a linear relationship between teachers’ allocations of time to teaching and to parents in the Dictator Game. The Spearman’s rho, the pairwise correlation and the linear regression with no clustering of errors return significance levels better than 5%. At the same time, the linear regression with errors clustered to account for possible interdependence within schools returns a significance level of 6.5%.

Finally, the linear relationship shown in **Figure 1** is explored in the summarizing tables below:

Dependant variable: Time allocation	
Constant	0.823
DG allocation	0.066
R-squared	0.009
Obs	487

Table 2: Summary of testing external validity of the DG by correlation

[5] Barr, A. and Zeitlin, A., 2010: “Dictator Games in the Lab and in nature: Evidence of external validity from Ugandan primary schools”

Table 3: Regression analysis of relationship between Dictator Game allocations to parents and time allocations to teaching

Dependent variable = Proportion of contracted time allocated to teaching					
	1a	2a	3a	4a	5a
Constant	0.823***	0.836***	1.001***	1.206***	1.212***
	(0.017)	(0.014)	(0.042)	(0.042)	(0.045)
Proportion allocated to parent in DG	0.066*	0.067**	0.067**	0.072**	0.166***
	(0.035)	(0.030)	(0.030)	(0.031)	(0.053)
In(adj(wealth$_i$-wealth$_{-i}$))$^{\#}$			-0.008***	-0.019***	-0.019***
			(0.002)	(0.002)	(0.002)
SMC Meetings * Allocation to parent in DG					-0.046**
					(0.022)
School fixed effects included	No	No	No	Yes	Yes
F-stat for school fixed effects				1.850***	1.890***
Observations	487	476	435	435	435

Notes: # ln(adj(wealth$_i$-wealth$_{-i}$)) is the natural log of the teacher's wealth minus the wealth of the average parents in their experimental session minus the within sample minimum of this difference plus 1; *** significant at 1% level; ** significant at 5% level; * significant at 10% level.

An R^2 near 1.0 indicates that a regression line fits the data well, while an R^2 closer to 0 indicates a regression line does not fit the data very well. It provides a measure of how well future outcomes are likely to be predicted by the model. In **Table 2** we can see a R^2 equal to 0.009, which is not close to 1 but to 0. Moreover, the coefficient in **Table 3** named "proportion allocated to parent in DG" is significant but very small, which means that the external validity is only *small*.

Therefore, this study provides weak evidence of local external validity. We wondered why this happens so we analyzed some differences between the natural decision-making context and the lab that might explain this weak correlation.

Differences between the natural decision-making context and the lab

Although the study of Barr and Zeitlin shows correlation and concludes that there actually is external validity in the Dictator Game they performed in their economic lab, we have identified some differences between the natural decision-making context and the laboratory environment that should have been taken into account by the researchers or at least that are worth mentioning to make this paper more complete.

First of all, in the natural decision-making context (being the Ugandan teacher's real life), teachers can coordinate their absences using or hiring a substitute, minimizing this way the impact on their students. At the same time, teachers may be called upon to cover for the absences of their colleagues so it is clear that absenteeism exerts a **negative externality** to the whole faculty.

In the DG experiment, though, decisions were made individually and there was no possibility of coordination, so negative externalities were indeed present but only to the hypothetically extent that the fact that a teacher refrained from going to work (which us would call an "egoistic" behavior) could cause a potential reputational harm to the whole faculty in the same experimental session. Moreover, the typical DG played in an economic laboratory is one-shot, whereas when played in the Ugandan study the time allocation decisions were repeated.

The next difference is that in nature, the teacher's time allocation decisions are observable to anyone in the community, including the SMC, who can choose between paying attention or not (and consequently doing something about it or not), but in the DG the teacher's allocation decisions are neither observed nor observable. Plus, teachers are hired to allocate a certain amount of their time to teaching and even though the contracts are not enforced they may (should) have an effect that is not replicated in the lab.

Finally, in the study DG is played with Ugandan Shillings, whereas the time allocation decisions involved time not money.

Finally, let us point out that this study does not take into account that human behavior may present significant differences between the lab and the outside world because human decisions are influenced not just by simple monetary calculations, but also by at least three other considerations:

1. The particular circumstances by which a decision is made (the lab environment)
2. The nature and extent to which one's actions are scrutinized by others
3. Self-selection of the individuals making the decisions.

In the last part of this paper we will try to assess how each of these factors influences decision making and the extent to which the environment constructed in the lab does or does not conform to real-world interactions on these various dimensions.

Model

We will use a model from Levitt and List (2006) in order to support our arguments and hypothesis regarding the potential factors that might influence individual decision-making. Neither Levitt and List nor us claim originality in the ideas modeled in this paper: indeed, starting with Smith in 1759, there has been a long list of economists who published papers and studies emphasizing that decisions can have an impact on individual utility that goes way beyond changes in wealth. In particular, we present a model that works under the assumption that utility is additively separable in the moral and wealth arguments. So, an individual *i* that takes a decision *a*:

$$U_i(a,v,n,s) = W_i(a,v) + M_i(a,v,n,s)$$

The function above[6] has to be read as follows: a utility-maximizing individual *i* is faced with a choice regarding a single action $a \in (0,1)$. The choice itself affects the agent's utility through two main channels: first, we find the effect on the individual's wealth (denoted W_i). The higher the stakes or monetary value of the game, which we denote *v*, the greater the decision's impact on W_i. Second, there is the non-monetary **moral cost or benefit** associated with action *i*, which we denote as M_i. If an individual is altruistic, for example, she or he will derive more or less utility from charitable contributions. This way, decisions that an individual views as anti-social, immoral, or contrary to her or his own identity may impose costs on the decision maker. In a dictator game, for example, keeping a greater share for oneself increases an individual's wealth, but doing so may cause the agent moral disutility (in the study we presented before, allocating a greater part of each dictator/teacher's money to themselves instead of allocating more to parents could cause them a moral cost, even though we think this would be unlikely to happen in reality as they *decide* not to go to

[6] Levitt, S., and List, J.A.; 2006: "What do Laboratory Experiments Tell Us About the Real World?"; University of Chicago and NBER.

work). This moral payoff might vary across culture, religions, or societies. We have focused on just three aspects of the moral determinant.

The first factor that influences the moral choice is the set of legal and social norms and the culture that drive behavior in a particular society, in our model denoted as n (should be read as social norms against an action a). Although there is not a financial externality supported by the recipient, there is indeed a potential moral costs associated with such behavior.

The second factor that we encounter is the financial externality that an action imposes on others. The greater the negative impact of an action is on others, the more negative the moral payoff M_i. The model considers the externality as an increasing function of the stake of the game v.

Finally, the third factor involves the moral concerns, which will be higher when the process by which a decision and final allocation are reached is emphasized or an individual's actions are more closely scrutinized: if the act is being televised (a great example is the worldwide famous prisoner's dilemma televised show: Golden Balls[7]) it is taking place in front of the agent's children or it is performed under the watchful eye of the researcher (as in the case of the Ugandan study). In the model, we denote the effect of scrutiny as s, with higher levels of scrutiny associated with greater moral costs. Four predictions can be derived once solved the simple decision problem:

First, because individuals follow different moral codes (that is, $M_i \neq M_j$ for individuals i and j), they will generally make different choices when faced with the same decision problem. Second, when the action that maximizes wealth has a moral cost associated to it, the agent will weakly (or not so weakly) deviate from the action towards one that imposes a lower moral cost. Third, the more important or restrictive

[7] Broadcasted in BBC June 2007-December 2009. Two people have to decide whether they want to split an amount of money or *steal* it and therefore keep all the money for him/herself leaving the other player with nothing. The dominant strategy for both players is, of course, to steal the money while trying to convince the other to choose "split", so the first player gets all the money. This, however, is generally seen as immoral and even more when it is being televised.

	Split	Steal
Split	50, 50	0, 100
Steal	100, 0	**0, 0**

is the social norm against the choice that maximizes wealth or the degree of scrutiny, the larger the deviation from that choice. In both cases, it is expected that the agent will trade-off morality and wealth until equilibrium is reached.

Finally, in situations that lack a moral component, for instance when investing in the stock market, the model goes back to the form of a standard wealth maximization problem.

This way, it is observed that in a simple \$5 dictator game a great number of players might transfer \$2.50 to their anonymous partner, but as soon as the stake rises to, for instance, \$500, it is expected an increase in the level of money transferred but not in an equivalent proportion. In such cases, the strategy "split the money equally" seems to be too costly to implement. We will go through empirical evidence later on.

Another factor that must be taken into account when talking about rises in the stakes is that as stakes rise, the moral penalty for violating a given norm will be greater: for example, people keenly disapprove shoplifting, but instead people are much more permissive, tolerant or forgiving of that crime than they are of bank robbery (and even more if there is violence involved). Similar reasoning works for stakes and scrutiny: an individual faces a greater utility loss of robbing a bank if his or her capture is broadcast in TV rather than merely recorded in his rap sheet. Moreover, relevant social norms and the amount of scrutiny are not necessarily exogenously determined, but they are usually subject to influence by those who will be affected by the choices an agent makes. For instance, churches use "open" rather than "closed" collection baskets in their masses in order to appeal to morality, shame and duty to get more money. This is consistent with the recognition of the importance of norms and scrutiny, as potential contributors can see the total amount already gathered and, more importantly, direct neighbors can witness each other's contributions.

As mentioned before, we will now examine some examples that constitute empirical evidence regarding each of the possible complications to extrapolating the experimental findings outside the lab reported in the model. What it is interesting is whether the depth and scope of such behaviors measured in the lab are shared widely among individuals in the field.

· **Scrutiny**

Scrutiny can take many dimensions, but in order to clarify and narrow down the discussion we will focus only on two dimensions, which we denote as "lab" effects and "non-anonymity" effects.

a) Lab effects

In a typical lab experiment, subjects *are* aware that their behavior is being monitored, recorded, and then scrutinized, thus, it is possible that behavior in the lab is more influenced by moral concerns and less aligned with wealth maximization than behavior in many naturally-occurring settings.

If we take a look at the study Bandiera et al. performed in 2010, we will find strong evidence of what we were just mentioning: they performed a study where they made use of personnel data from a leading United Kingdom based fruit farm, and they found that behavior is consistent with a model of social preferences when workers can be monitored, but when workers cannot be monitored, pro-social behaviors disappear. A clearer example of the fact that being watched modifies keenly human behavior is provided with the study Benz and Meier published in 2005. They found some evidence of correlation across situations (that is, evidence of external validity; as in the Ugandan study showed in this paper, they found evidence strong enough to be able to say that the same behavior that occurs in the lab can be found in nature), but they found that subjects who have never contributed in the past to the charities gave 75% of their endowment to the charity in the lab experiment. Similarly, those who never gave to the charities subsequent to the lab experiment gave more than 50% of their experimental endowment to the charities in the lab experiment.

b) Non-anonymity effects

We can define the non-anonymity effects as how changes in the degree of confidentiality may influence behavior. In a typical lab experiment subjects are anonymous in relation to other subjects, but the identity of the subject can be easily linked to individual choices by the researcher. For instance, Haley and Fessler (2005) found that the amount given in a dictator game significantly increased when a pair of eyes was shown on the computer screen with which the dictator made his/her allocation. This simple manipulation increased the proportion of nonzero givers from

55% in the control (regular) treatment to 88% in the treatment that included the pair of eyes.

· Stakes

It is predicted that in games that have both a morality and wealth component, financial concerns and stakes have a directly proportional relation: as stakes increase, financial concerns also increase. In the second mover player in ultimatum games[8], for example, the acceptance rate is increasing in the amount offered, conditional on the share offered (that is, a $1 offer in a $5 game is rejected more often than a $100 offer in a $500 game). We find evidence in the study Slonim and Roth (1998) published: they found that in each range of offers below 50% the acceptance rate goes up as the level of stakes increase (from 60 to 1500 slovak Koruna, respectively).

In Carpenter et al. (2005a), similarly, it is shown that an increase in stakes from $10 to $100 in a dictator game caused the median offer to drop from 40% to 20% of the endowment.

Conclusions and discussion

This paper starts by testing by correlation the external validity of a simple one-shot Dictator Game as a laboratory analogue for a specific, naturally occurring and policy-relevant decision-making context and explore several possible factors that explain why the correlation was not perfect.

The naturally occurring decision-making context was the one in which primary school teachers in Uganda decide how much of their time specified in their contract they actually allocate to teaching. Researchers Barr and Zeitlin ran a Dictator Game in school classrooms where teachers were the “dictators” and pupil’s parents were the “passive recipient”, and they used the local currency as stake. The Dictator Game among all the Experimental Economics experiments available was selected as the laboratory analogue for this context for two reasons: it generated an easy way for

[8] The **ultimatum game** is a game often played in economic experiments is a two-players game in which the agents interact to decide how to divide the endowment that is given to them. The first player proposes how to divide the sum between the two players, and the second player can either accept or reject this proposal. If the second player rejects, neither player receives anything. If the second player accepts, the money is split according to the proposal. This is a one-shot game so reciprocation is not an issue.

readily interprete measure of teachers' intrinsic motivations and it was also a good match to a stylized characterization of the teachers' current decision-making context.

Several tests showed that the teachers' allocations to parents in the Dictator Game were positively correlated with their time allocations to teaching in reality. That is, the more endowment a certain teacher allocated to the parents' picture when playing the DG, the more time this particular teacher actually allocated to teaching in reality (data gotten from surveys ran by the same researchers). Even though they found the positive correlation we just explained, it was very weak, as can be seen in the p-values table (**Table 1**) and the summarizing table with tested correlation (**Table 3**) provided before, where it can be seen a small correlation coefficient (only significant at the 10% level).

Precisely because of this weak correlation we decided to dig deeper into the possible factors that may influence the external validity of any economic experiment. Arguing that experiments may not always yield results that are robustly generalizable, we state that the choices individuals make depend not just on financial implications, but also on the nature and degree of others' scrutiny and the particular context in which a decision is embedded. Because laboratory and naturally-occurring environments systematically differ on any of these dimensions, the results obtained inside and outside the lab need not correspond.

The argument provided is that laboratory experiments usually present a special type of scrutiny: a context that places extreme emphasis on the process by which decisions and allocations are reached. However, many real-world markets present a different type of scrutiny, little focus on process, and very different forms of self-selection of participants.

There is still an important role for traditional laboratory experiments in economics, but maybe this role is more limited than experimentalists might subscribe. It is clear that experiments can provide a crucial first understanding of qualitative effects, they suggest underlying mechanisms that might work or appear when some data patterns are observed and they sure provide insights into what can happen.

In the discussion we would like to present three important conclusions regarding research design and interpretation. First of all, combining laboratory analysis with a model of decision-making expands the potential role and power of lab experiments. Anticipating the biases that typically appear in the lab would help designing new and improved experiments that would minimize such biases. Moreover, knowing the magnitude of any biases included by the lab, one can extract useful information from a study, even if the results cannot be extrapolated outside the lab.

Second, by focusing on qualitative rather than quantitative insights much can be learned: adopting experimental designs that recognize the potential weaknesses of the laboratory constructions would increase the usefulness of lab studies.

Finally, let us point out that we believe that the strong dichotomy usually drawn between data generated in laboratory experiments and data collected in nature is a false one. The same concerns arise in both settings regarding their generalizability outside of the immediate application, circumstances, and treated population. Each approach has different advantages and disadvantages, and thus a combination of the two would definitely provide more insight than either in isolation.

References:

Bandiera, O., Barankay, I., Rasul, I., 2010: "Team incentives: Evidence from a Firm level experiment"; London School of Economics.

Barr, A., and Zeitlin, A., 2010: "Dictator Games in Lab and in nature: Evidence of external validity from Ugandan primary schools"; University of Oxford.

Benz, M., and Meier, S., 2005: "Do we behave in Experiments as in the field? Evidence from donations"; Federal Reserve Bank of Boston, Research Center for Behavioral Economics and Decision-Making.

Campbell, D. T., and Stanley J.C., 1963: "Experimental and Quasi-Experimental Designs for Research"; Boston: Houston Mifflin.

Carpenter, J., Holmes, J., Matthews, P. H., 2008: "Charity auctions: a field experiment": *The Economic Journal*, vol.118 92-113

Chaudhury, N., Hammer, J., Kremer, M., Muralidharan, K., And Rogers, F., 2006: "Missing in Action: Teacher and Health Worker Absence in Developing Countries." *Journal of Economic Perspectives*, 20(1): 91-116.

Cross, J., 1980: "Some comments on the Papers by Kagel and Battalio and by Smith", in J. Kmenta and J. Ramsey (eds.), *Evaluation of Econometric Models*; NYU Press: 403-406.

Feynman, R.P., Leighton R.B. and Sands M., 1963: "The Feynman Lectures on Physics"; Reading, MA: Addison-Wesley Publishing Company.

Guala, F., 2002: "On the scope of Experiments in Economics: Comments on Siakantaris"; *Cambridge Journal of Economics*, 26: 261-267.

Guala, F., 2008: "Experimental Economics, History of", *The new Palgrave Dictionary of Economics*, Vol. 3, edited by S. Durlauf and L. Blume. London: Palgrave-MacMillan (2008), pp. 152-156.

Haley, K., Fessler, D., 2005: "Nobody's watching? Subtle cues affect generosity in an anonymous economic game", *Evolution and Human Behavior*, Elsevier, 245-256

Kahneman, D., Smith, V., 2002: "Foundations of Behavioral and Experimental Economics"; *Advanced information on the prize in Economic Sciences*, The Royal Swedish Academy of Sciences.

Levitt, S., and List, J.A., 2006: "What do Laboratory Experiments Tell Us About the Real World?"; University of Chicago and NBER.

Loweinstein, G., 1999: "Experimental Economics from the vantage-point if Behavioral Economics"; *The Economic Journal*, 109: F25-F34.

Martel Garcia, F., Wantchekon, L., 2009: "Theory, External Validity and Experimental Inference: Some Conjectures"; New York University.

Plott, C.R., 1982: "Industrial Organization Theory and Experimental Economics"; *Journal of Economic Literature*, 20: 1485-1527.

Samuelson, P. and Nordhaus, W., 1985. Economics, McGraw-Hill.

Schram, A., 2005: "Artificiality: the tension between internal and external validity in Economic Experiments"; CREED, University of Amsterdam.

Settle, Thomas B., 1961: "An experiment in the History of Science"; *Science*, vol. 133, pp. 19-23.